10/568440

IAP20 Rec'd PCT/PTO 14 FEB 2006

SPECIFICATION

SESSION RELAY APPARATUS AND RELAYING METHOD

TECHNICAL FIELD

[0001]

The present invention relates to a session relay apparatus and a session relaying method used therefor, and more particularly, to an apparatus for relaying data during a TCP (Transmission Control Protocol) session.

BACKGROUND ART

[0002]

Generally, in communication applications, a communication session is established between a transmission terminal and a reception terminal, and communication is made on the established session. However, when a propagation delay time is very long between the transmission terminal and the reception terminal, or when communication is made across networks which differ in characteristics such as wired and wireless ones, the communication throughput degrades between the transmission terminal and the reception terminal.

[0003]

As methods for solving this problem, there exist systems as disclosed in the following Documents 1 - 3. In this system, instead of making communication between a transmission terminal and a reception terminal in a single session, a relay apparatus is

installed between the transmission terminal and the reception terminal. Then, communication is made by relaying data in two sessions, i.e., a session from the transmission terminal to the relay apparatus and a session from the relay apparatus to the reception terminal.

[0004]

Document 1: JP-A-11-252179

Document 2: JP-A-2002-281104

Document 3: Bakre and B.R.Badrinath, "I-TCP; Indirect TCP for Mobile Host", Department of Computer Science Rutgers University, DSC-TR-314, 1994 (http://www.it.iibt.ac.in/it644/papers/i-tcp.pdf)

A relay apparatus that has a large transmission buffer in order to prevent the throughput from degrading even if a propagation delay time is very long is known to be effective. However, when a large transmission buffer is contained, packets are delivered in a burst manner particularly when the TCP session starts slowly which causes congestion on a network, resulting in a lower throughput.

[0005]

As a method for solving such a problem, there is a method which adjusts intervals at which packets are delivered in such a manner that the packets are not delivered in a burst manner from a TCP session, as disclosed in the following Document 4:

[0006]

Document 4: A. Aagarwal, S.Savage, and T.Anderson,

"Understanding the Performance of TCP Pacing", in proceedings of IEEE INFOCOM' 2000

Also, as another method, output packets from a TCP session are placed in a queue and controlled to be delivered by a scheduler, thereby making it possible to deliver packets in an arbitrary bandwidth, to limit the throughput of another TCP session, and to improve the throughput of a particular session as well.

[0007]

For relaying a session, the most frequent problem is that a processing load is generally high in processing of the TCP session so that fast relay processing is difficult. Also, when an adjustment is made for a packet delivery interval from the TCP session, the processing load becomes even higher.

[0008]

As a system for speeding up the processing of the TCP session, there exist systems as described below. As a first system, there is a system for reducing the processing load on a relay apparatus by using a zero copy method for data transfer between the kernel of OS (Operating System) and an application program. In this first system, when data is passed between a kernel program for actually performing data transmission/reception processing and an application program for performing data relay processing, virtual data movement is performed through page mapping without performing a physical data copy.

[0009]

As a second system, there is a system which only

conducts retransmission control when relay processing is performed. In this second system, a session is not terminated in a relay apparatus, but one session is set between a transmitting and a reception terminal. Then, in the relay apparatus, when a packet received from the transmission terminal is delivered to the reception terminal, the packet is not only transferred, but is also saved in the relay apparatus.

[0010]

The relay apparatus monitors an ACK (acknowledgement) packet returned from the reception terminal to the transmission terminal, delivers the saved packet to the reception terminal upon detection of a discarded packet between the relay apparatus and the reception terminal, to prevent a retransmission of the packet from the transmission terminal and to prevent a lower throughput between the transmission terminal and the reception terminal due to the discarded packet.

[0011]

In this second system, relay processing is not performed between sessions, the retransmission processing alone is performed, so that the load on the relay apparatus can be reduced, as compared with the case where relay processing is performed.

[0012]

As a third system, there is a system which performs relay processing only for a session which requires a relay, without performing the relay processing for a session which does not require a relay, thereby reducing the processing load on a relay

apparatus. For example, as an illustration of this system, there is a system which measures the throughput of a session, and performs the relay processing only for selected sessions which allow for an improved throughput, if the relay processing is performed therefor, as described in the following Document 5.

[0013]

Document 5: JP-A-11-112576

Also, as the aforementioned system, there is a system which does not perform relay processing when an HTTP (Hyper Text Transfer Protocol) request is transmitted, but performs the relay processing only when data is transferred, as disclosed in the following Document 6.

[0014]

Document 6: JP-A-2002-312261

In the aforementioned conventional session relay systems, the first system has the problem that the processing load is also high besides the data copies. When delivery control is conducted in accordance with a TCP pacing system or a scheduler, a load for packet delivery control is added, in addition, to a TCP processing load, resulting in a higher overall processing load.

[0015]

Also, when relay processing is performed for an upper layer protocol such as, for example, iSCSI (internet Small Computer System Interface), as well as for a relay of TCP sessions, packet delivery control is further added on the basis of congestion control on an upper layer, resulting in a yet higher processing load.

Particularly, when there are a large number of sessions to be relayed, a higher processing load is required for conducting delivery control for packets between sessions.

[0016]

Further, in conventional session relay systems, the first system has a problem in that the packet length is limited. To move virtual data between a TCP layer and an application in accordance with page mapping, the page size of a CPU (central processing unit) must match the packet length.

[0017]

Generally, a session relay apparatus exists independently of transmitting and reception terminals, and the packet length cannot be previously assumed for use by the transmission terminals and reception terminals, so that no reduction in processing load can be expected by eliminating copies for those transmission and reception terminals which employ a packet length different from the page size of a bandwidth control apparatus.

[0018]

On the other hand, in the conventional session relay systems, the second system has a problem in that it cannot improve a reduction in the throughput due to factors other than discarded packets. When session relay processing is performed, the session relay apparatus immediately returns an ACK packet to a transmission terminal in response to packets received from the transmission terminal, that area intended for confirming the reception. However, in this second system, a reception terminal

simply returns an ACK packet to a transmission terminal, whereas the session relay apparatus itself does not return an ACK packet. For this reason, the second system provides a smaller throughput improving effect in an environment in which a propagation delay is large.

[0019]

Also, in the conventional session relay systems, the third system has a problem in that the processing load is not reduced when there are a large number of sessions which must be relayed.

DISCLOSURE OF THE INVENTION

[0020]

It is therefore an object of the present invention to provide a session relay apparatus which is capable of solving the problems mentioned above and performing relay processing between sessions at high speeds even when there are a large number of sessions to be relayed, and packet delivery control is conducted, and a session relaying method used therefor.

[0021]

A session relay apparatus according to the present invention is a session relay apparatus for performing session relay processing including congestion control processing and packet delivery control processing on a plurality of layers, wherein each of the plurality of layers only creates the congestion control information, and the packet delivery control processing is concentrated in a scheduler on an IP (Internet Protocol) layer.

[0022]

Another session relay apparatus according to the present invention is a session relay apparatus for realizing communication between a reception terminal and a transmission terminal by relaying data between a session to the transmission terminal and a session to the reception terminal, comprising:

reception session processing means for receiving data from the session to the transmission terminal, transmission session processing means for transmitting data to the session to the reception terminal, transmission buffer for temporarily storing data delivered to the transmission terminal, a packet scheduler for controlling a packet delivery from the transmission buffer, and delivery control means for controlling the delivery of data stored in the transmission buffer in response to the control of the packet scheduler,

wherein the transmission session processing means calculates the amount of data which is permitted to be delivered on the layer, and the packet scheduler controls the packet delivery based thereon.

[0023]

Another session relay apparatus according to the present invention is a session relay apparatus for realizing communication between a transmission terminal and a reception terminal by relaying data between a session to the transmission terminal and a session to the reception terminal, comprising:

reception session processing means provided in correspondence to a plurality of layers for receiving data from the

session to the transmission terminal, transmission session processing means provided in correspondence with the plurality of layers for transmitting data to the session to the reception terminal, a transmission buffer for temporarily storing data delivered to the transmission terminal, and a packet scheduler for controlling the delivery of packets from the transmission buffer,

wherein each of the transmission session control means calculates the amount of data permitted to be delivered on an associated layer, and the packet scheduler controls the packet delivery based on the amount of data permitted in common on all of the plurality of layers.

[0024]

A session relaying method according to the present invention is a session relaying method for a session relay apparatus for performing session relay processing including congestion control processing and packet delivery control processing on a plurality of layers, wherein each of the plurality of layers only creates congestion control information, and the packet delivery control processing is concentrated in a scheduler on an IP (Internet Protocol) layer.

[0025]

Another session relaying method according to the present invention is a session relaying method for a session relay apparatus for realizing communication between a reception terminal and a transmission terminal by relaying data between a session to the transmission terminal and a session to the reception terminal,

comprising, on the session relay apparatus side, reception session processing of receiving data from the session to the transmission terminal, transmission session processing for transmitting data to the session to the reception terminal, processing for temporarily storing data delivered to the transmission terminal in a transmission buffer, processing for controlling a packet delivery from the transmission buffer in a packet scheduler, and processing for controlling the delivery of data stored in the transmission buffer in response to the control of the packet scheduler in delivery control means, wherein the transmission session processing calculates the amount of data which is permitted to be delivered on the layer, and the packet scheduler controls the packet delivery based thereon.

[0026]

Another session relaying method according to the present invention is a session relaying method for a session relay apparatus for realizing communication between a transmission terminal and a reception terminal by relaying data between a session to the transmission terminal and a session to the reception terminal, comprising on the session relay apparatus side, reception session processing for receiving data from the session to the transmission terminal in each of the plurality of layers, transmission session processing for transmitting data to the session to the reception terminal in each of the plurality of layers, processing for temporarily storing data delivered to the transmission terminal in a transmission buffer, and processing for controlling the delivery of packets from the transmission buffer in a packet scheduler, wherein each of the

transmission session control processing calculates the amount of transmission permitted data on an associated layer, and the packet scheduler controls the packet delivery based on the amount of data permitted in common on all of the plurality of layers.

[0027]

Specifically, the session relay apparatus of the present invention does not control the packet delivery on the TCP (Transmission Control Protocol) layer, but only generates control information for the packet delivery on the TCP layer, and controls the packet delivery using the packet scheduler on the IP (Internet Protocol) layer.

[0028]

Also, in the session relay apparatus of the present invention, control information alone is created on an higher-rank layer even in regard to congestion control in a higher-rank layer protocol such as iSCSI (internet Small Computer System Interface), and uses the packet scheduler on the IP layer for actual packet delivery control.

[0029]

In this way, in the session relay apparatus of the present invention, since the data delivery control on a plurality of layers can be integrated, it is possible to reduce the processing load relating to the packet delivery.

[0030]

Further, in the session relay apparatus of the present invention, since no data is copied to an application, there is no need

for movements of data involved in page mapping. Specifically, a received packet is directly stored in the transmission buffer of the IP layer without being stored in a reception buffer of the TCP layer or other higher-rank layers, or an application or without being stored in a transmission buffer of the TCP layer or other higher-rank layers, or an application. For this reason, in the session relay apparatus of the present invention, no data movement occurs within the session relay apparatus, so that there is no need for movement of data involved in page mapping.

[0031]

Furthermore, in the session relay apparatus of the present invention, since the relay processing is not performed only for packet retransmission control, but a session is once completely terminated to perform complete relay processing, it is possible to improve a degraded throughput due to factors other than discarded packets.

[0032]

In the session relay apparatus of the present invention, the relay processing between sessions can be performed at higher speeds, so that even when there is a large number of sessions to be relayed, the processing can be performed at high speeds.

BRIEF DESCRIPTION OF THE DRAWINGS

[0033]

[Fig. 1]

Fig. 1 is a block diagram illustrating the configuration of a transmission system which includes a session relay apparatus

according to a first embodiment of the present invention.

[Fig. 2]

Fig. 2 is a block diagram illustrating the configuration of the session relay apparatus according to the first embodiment of the present invention.

[Fig. 3]

Fig. 3 is a block diagram illustrating the configuration of a packet scheduler in Fig. 2.

[Fig. 4]

Fig. 4 is a flow chart illustrating the operation of the session relay apparatus according to the first embodiment of the present invention.

[Fig. 5]

Fig. 5 is a flow chart illustrating the operation of the session relay apparatus according to the first embodiment of the present invention.

[Fig. 6]

Fig. 6 is a flow chart illustrating the operation of the packet scheduler illustrated in Fig. 3.

[Fig. 7]

Fig. 7 is a schematic diagram showing the number of transmission waiting bytes.

[Fig. 8]

Fig. 8 is a block diagram illustrating the configuration of a packet scheduler according to a second embodiment of the present invention.

[Fig. 9]

Fig. 9 is a block diagram illustrating the configuration of a session relay apparatus according to a third embodiment of the present invention.

[Fig. 10]

Fig. 10 is a block diagram illustrating the configuration of a session relay apparatus according to a fourth embodiment of the present invention.

[Fig. 11]

Fig. 11 is a schematic diagram describing the number of transmission waiting bytes in the fourth embodiment of the present invention.

[Fig. 12]

Fig. 12 is a block diagram illustrating the configuration of a session relay apparatus according to a fifth embodiment of the present invention.

[Fig. 13]

Fig. 13 is a block diagram illustrating a flow of data between the session relay apparatus (transmission terminal) and the session relay apparatus (reception terminal) according to the fifth embodiment of the present invention.

BEST MODE FOR CARRYING OUT THE INVENTION

[0034]

Next, embodiments of the present invention will be described with reference to the drawings.

(First Embodiment)

Fig. 1 is a block diagram illustrating the configuration of a transmission system which includes session relay apparatus 1 according to a first embodiment of the present invention. In Fig. 1, session relay apparatus 1 according to this embodiment comprises session identification unit 11, session relay units 12-1 - 12-N, and delivery control unit 14, and is connected to reception terminal 2 and transmission terminal 3.

[0035]

First, when data is sent from transmission terminal 3 to reception terminal 2, a data packet from transmission terminal 3 is processed by a reception session processing unit (not shown) of session relay unit 12-1, and as a result, an ACK (acknowledgement) packet is returned to transmission terminal 3.

[0036]

Data received by the reception session processing unit of session relay unit 12-1 is sent to a transmission session processing unit (not shown) of session relay unit 12-1, and a data packet is transmitted from here to reception terminal 2. On the other hand, the ACK packet returned by reception terminal 2 is processed by the transmission session processing unit of session relay unit 12-1.

[0037]

Likewise, when data is sent from reception terminal 2 to transmission terminal 3, a data packet from reception terminal 2 is processed by a reception session processing unit (not shown) of session relay unit 12-2, and as a result, an ACK packet is returned to reception terminal 2.

[0038]

The data received by the reception session processing unit of session relay unit 12-2 is sent to a transmission session processing unit (not shown) of session relay unit 12-2, and a data packet is transmitted from here to transmission terminal 3. On the other hand, the ACK packet returned by transmission terminal 3 is processed by the transmission session processing unit (not shown) of session relay unit 12-2.

[0039]

Fig. 2 is a block diagram illustrating the configuration of the session relay apparatus according to the first embodiment of the present invention. In Fig. 2, session relay apparatus 1 comprises session identification unit 11, session relay units 12-1 - 12-N, packet scheduler 13, and delivery control unit 14.

[0040]

Session identification unit 11 determines a session to which an incoming packet belongs. Session relay unit 12-1 - 12-N relays packets between a session with transmission terminal 3 and a session with reception terminal 2. Packet scheduler 13 controls packets outputs from each of session relay units 12-1 - 12-N. Delivery control unit 14 delivers packets from each session relay unit 12-1 - 12-N based on instructions from packet scheduler 13.

[0041]

Session relay unit 12-1 in turn comprises transmission session processing unit 121-1 for processing a session to transmit data to reception terminal 2; transmission buffer 122-1 for storing

received data until the end of a transmission; and reception session processing unit 123-1 for processing a session to receive data from transmission terminal 3. Though not shown, the configuration of session relay units 12-2 - 12-N is the same as the configuration of session relay unit 12-1 described above.

[0042]

Fig. 3 is a block diagram illustrating the configuration of packet scheduler 13 in Fig. 2. In Fig. 3, packet scheduler 13 comprises list distribution unit 131, state update unit 132, state variable saving unit 133, reset control unit 134, transmission available list 135, and transmission waiting list 136.

[0043]

Transmission available list 135 holds identifiers of sessions in which packets can be delivered, while transmission waiting list 136 holds identifiers of sessions which are waiting for transmission. List distribution unit 13 distributes an identifier of a data packet session or an identifier of a session for which an ACK packet has been received to transmission available list 135 or to transmission waiting list 136.

[0044]

State update unit 132 retrieves the identifier of a session from transmission available list 135 for notification to delivery control unit 14, and updates the state of the session. State variable saving unit 133 holds the state of each session, while reset control unit 134 moves the identifier of a session managed by transmission waiting list 136 to transmission available list 135.

[0045]

In a TCP (Transmission Control Protocol) session, bidirectional communications are generally made between transmission terminal 3 and reception terminal 2. Thus, assume in this embodiment that two session relay units are used for a set of transmission terminal 3 and reception terminal 2, and corresponding session relay units are used respectively for data communications in the respective directions.

[0046]

Therefore, session relay units 12-1 - 12-N are provided two by two for a plurality of sets of transmission terminals 3 and reception terminals 2, and each of session relay units 12-1 - 12-N performs processing for relaying data from a session of transmission terminal 3 to a session of corresponding reception terminal 2, or from a session of reception terminal 2 to a session of corresponding transmission terminal 3.

[0047]

In the TCP session, a data packet in a certain direction and an ACK packet in a direction opposite thereto can be integrated onto a single packet (piggy back of ACK), but a description on such an operation is omitted in this embodiment to simplify the description.

[0048]

Figs. 4 and 5 are flow charts illustrating the operation of session relay apparatus 1 according to the first embodiment of the present invention, and Fig. 6 is a flow chart illustrating the operation

of packet scheduler 13 illustrated in Fig. 3. The operation of session relay apparatus 1 according to the first embodiment of the present invention will be described with reference to these Figs. 1 - 6.

[0049]

When a packet is fed to session relay apparatus 1 according to this embodiment (step S1 in Fig. 4), session identification unit 11 references a header of the packet, and determines a session to which the packet belongs, based on a source IP (Internet Protocol) address, a destination IP address, a fourth-layer protocol number, a source fourth-layer port number, a destination fourth-layer port number, and the like (step S2 in Fig. 4).

[0050]

When the packet is a data packet (step S3 in Fig. 4), session identification unit 11 passes the packet to reception session processing unit 123-1 - 123-N of corresponding session relay unit 12-1 - 12-N (reception session processing units 123-2 - 123-N are not shown) (step S4 in Fig. 4).

[0051]

On the other hand, when the packet is an ACK packet (step S3 in Fig. 4), session identification unit 11 passes the packet to transmission session processing unit 121-1 - 121-N of corresponding session relay unit 12-1 - 12-N (transmission session processing units 121-2 - 121-N are not shown) (step S10 in Fig. 4).

[0052]

Reception session processing unit 123-1 - 123-N sorts

data packets applied thereto in the order of sequence numbers for storage in transmission buffer 122-1 - 122-N (transmission buffers 122-2 - 122-N are not shown) (step S5 in Fig. 4). Also, when a received data packet has a correct sequence number (step S6 in Fig. 4), i.e., when the sequence number is continuous to the sequence number at the end of data which have been previously received in sequence, reception session processing unit 123-1 - 123-N returns an ACK packet through delivery control unit 14 for notifying transmission terminal 3 of the acknowledgement of data reception and an advertisement window size (step S7 in Fig. 4).

[0053]

Further, reception session processing unit 123-1 - 123-N returns an ACK packet (duplicate ACK packet) based on the sequence number of the last one of sequentially received packets, and notifies transmission terminal 3 of unarrival of packets.

[0054]

Since the foregoing processing is described in detail in TCP/IP Illustrated, Volume 1: The Protocols, Addison-Wesley, 1994, ISBN 0-201-63346-9 (hereinafter called "Document 7"), a description thereon will not be made in detail.

[0055]

After an ACK packet is generated, it is immediately delivered from delivery control unit 14, or stored in transmission buffer 122-1 - 122-N of the corresponding session in the opposite direction and delivered together with data packets for the session in the opposite direction in response to an instruction from packet

scheduler 13.

[0056]

Also, reception session processing unit 123-1 - 123-N notifies packet scheduler 13 of the sequence number of the last one of sequentially received data (step S8 in Fig. 4).

[0057]

Transmission session processing unit 121-1 - 121-N changes the congestion window size based on the applied ACK packet (step S9 in Fig. 4), and erases data, the reception of which has been confirmed, from transmission buffer 122-1 - 122-N if the ACK packet is not a duplicate ACK (steps S12, S13 in Fig. 5), or retransmits data as required if it is a duplicate ACK (steps S12, S11 in Fig. 5). Since this processing is also described in detail in the aforementioned Document 7, a description thereon is not made in detail.

[0058]

Transmission session processing unit 121-1 - 121-N notifies packet scheduler 13 of the advertisement window described in the received ACK packet and in the updated congestion window (step S15 in Fig. 5). Also, when a retransmission times out (step S14 in Fig. 5), transmission session processing unit 121-1 - 121-N also notifies the last sequence number of packets, the reception of which has been confirmed (step S16 in Fig. 5).

[0059]

In this embodiment, packets are delivered only when the delivery of packets is instructed from packet scheduler 13, so that

when a duplicate ACK is received, packets are not immediately retransmitted, but the sequence numbers of packets to be retransmitted are simply stored as packets to be delivered next.

[0060]

The delivery of packet from transmission buffer 122-1 - 122-N is performed on the basis of an instruction from packet scheduler 13. When the delivery of packet is instructed from packet scheduler 13, delivery control unit 14 retrieves one packet from transmission buffer 122-1 - 122-N onto an outgoing line, and notifies packet scheduler 13 of the packet length of the delivered packet.

[0061]

A packet delivered from transmission buffer 122-1 - 122-N is a packet, the sequence number of which has been stored for performing retransmission processing, or the packet having the smallest sequence number of untransmitted packets.

[0062]

When a session of interest is not a TCP session, reception session processing unit 123-1 -123-N simply stores received packets in transmission buffer 122-1 - 122-N in the order in which they have arrived. Transmission session processing unit 121-1 - 121-N does not receive the ACK packet but notifies packet scheduler 13 of the queue length of transmission buffer 122-1 - 122-N as the advertisement window and congestion window, thereby continuously requesting packet scheduler 13 to deliver packets as long as packets are stored in transmission buffer 122-1 -

122-N.

[0063]

Packet scheduler 13 holds three parameters, an assignment weight, the number of transmissible bytes, and the number of transmission waiting bytes in state variable saving unit 133 on a session-by-session basis.

[0064]

The assignment weight is a weight assigned to the session. Packet scheduler 13 resets at a fixed period or each time there is no longer a session in which packets can be transmitted (steps S26, S27 in Fig. 6), and the number of bytes which can be transmitted during one reset period is the assignment weight.

[0065]

The number of transmissible bytes is the number of bytes which can be transmitted from a current time to the next reset, and has an initial value equal to the assignment weight (step S21 in Fig. 6), and is decremented by the packet length each time a packet is delivered (steps S22, 23 in Fig. 6).

[0066]

The number of transmission waiting bytes is a value derived by subtracting a sequence number already transmitted by packet scheduler 13 from a sequence number which is permitted by the TCP layer to be transmitted in the case of the TCP session, and represented by min(the sequence number of the last one of sequentially received data +1, the advertisement window indicated by the reception terminal, the congestion window of the session) -

the transmitted sequence number. For the case other than the TCP session, the number of transmission waiting bytes is a queue length in transmission buffer 122-1 - 122-N.

[0067]

Fig. 7 is a schematic diagram showing the number of transmission waiting bytes. In Fig. 7, the identifier of a session in which a packet can be delivered, i.e., a session which has the number of transmissible bytes and the number of transmission waiting bytes both equal to one or MSS (Maximum Segment Size) or more, is managed by transmission available list 135 (steps S24, S25 in Fig. 6), while the identifier of a session in which a packet can be delivered after the number of transmissible bytes is reset, i.e., a session which has the number of transmission waiting bytes equal to one or MSS or more but has the number of transmissible bytes equal to one or less than MSS, is managed by transmission waiting list 136 (steps S27 - S29 in Fig. 6).

[0068]

Next, the operation of packet scheduler 13 will be described with reference to Fig. 3. List distribution unit 131 updates the number of transmission waiting bytes in a manner shown in Fig. 7 when it receives the sequence number of the last one of packets sequentially received from transmission terminal 3, an advertisement window received from reception terminal 2, and an updated congestion window from session relay unit 12-1 - 12-N.

[0069]

Also, when a retransmission timer times out in delivery

control unit 14, list distribution unit 131 also returns a transmitted sequence number to a reception confirmed sequence number, and updates the number of transmission waiting bytes. When the number of transmission waiting bytes is one or less than MSS before the update, the identifier of the session is not managed by transmission available list 135 or transmission waiting list 136, so that list distribution unit 131 newly stores the identifier of the session in transmission available list 135 or transmission waiting list 136 based on the updated number of transmission waiting bytes and number of transmissible bytes.

[0070]

State update unit 132 retrieves one identifier of a transmission available session from the top of transmission available list 135, and notifies delivery control unit 14 of this to deliver a packet. Subsequently, state update unit 132 adds the length of the delivered packet to the transmitted sequence number, and then updates the number of transmission waiting bytes, and subtracts the length of the delivered packet from the number of transmissible bytes.

[0071]

State update unit 132 again stores the identifier of the session in transmission available list 135 or transmission waiting list 136 in accordance with the number of transmission waiting bytes and the number of transmissible bytes. Specifically, state update unit 132 stores the identifier in transmission available list 135 when the number of transmission waiting bytes and the number of

transmissible bytes are both one or MSS or more, and stores the identifier in transmission waiting list 136 when the number of transmission waiting bytes is one or MSS or more but the number of transmissible bytes is one or less than MSS.

[0072]

When transmission available list 135 becomes empty, state update unit 132 resets the numbers of transmissible bytes of all sessions, i.e., adds the assignment weight to the number of transmissible bytes. If the number of transmissible bytes is equal to or more than the assignment weight, the number of transmissible bytes is made equal to the value of the assignment weight [number of transmissible bytes = min(number of transmissible bytes + assignment weight, assignment weight)].

[0073]

Since this processing causes all sessions in a transmission waiting state to change to a transmission available state, sessions managed by transmission waiting list 136 are all moved to transmission available list 135.

[0074]

If the number of transmissible bytes is not one or less than MSS, and if there are data waiting for transmission in transmission buffer 122-1 - 122-N associated with the session before performing this reset processing, this is assumed to be a state in which there is a room in an output bandwidth from the session, but the transmission is stopped by the TCP control in transmission session processing unit 121-1 - 121-N.

[0075]

In this event, in order to keep the bandwidth in preparation for a future resumption of transmission, even if the sum of the number of transmissible bytes and the assignment weight exceeds the assignment weight, the number of transmissible bytes is kept up to a certain value [number of transmissible bytes = min(number of transmissible bytes + assignment wait, upper limit value for number of transmissible bytes)].

[0076]

As described above, in this embodiment, a relayed packet moves only when it is stored in transmission buffer 122-1 - 122-N upon reception and when it is retrieved from transmission buffer 122-1 - 122-N upon transmission, so that the data movement entails a small overhead. Also, since the packet delivery is controlled by packet scheduler 13 which uses information from the TCP layer, packets can be delivered while the bandwidth control is conducted with a reduced processing load than when packets are directly delivered from transmission session processing unit 121-1 - 121-N.

(Second Embodiment)

The configuration of a session relay apparatus according to a second embodiment of the present invention is similar to the configuration of session relay apparatus 1 according to the first embodiment of the present invention illustrated in Fig. 2.

[0077]

Fig. 8 is a block diagram illustrating the configuration of a

packet scheduler according to the second embodiment of the present invention. In Fig. 8, packet scheduler 16 according to the second embodiment of the present invention is similar in configuration to the first embodiment of the present invention illustrated in Fig. 3 except for the addition of assignment weight changing unit 161 and control parameter changing unit 162, and the same components are designated the same reference numerals. Also, the operation of the same components is similar to the first embodiment of the present invention.

[0078]

The operation of packet scheduler 16 according to the second embodiment of the present invention will be described with reference to Fig. 8. Here, a description will be given only of aspects different from packet scheduler 13 according to the first embodiment of the present invention.

[0079]

In the session relay apparatus according to this embodiment, control parameter changing unit 162 in packet scheduler 16 dynamically changes the value of the TCP control parameter for the session in accordance with a bandwidth which is set as the assignment weight. Specifically, control parameter changing unit 162 determines that the delivery of data is limited by the TCP control when the number of transmissible bytes of the session is not larger than a previously set value that is larger than the assignment weight, and changes the value of the TCP control parameter for the session such that the session can deliver data

over a wider bandwidth.

[0080]

However, when even a change in the TCP control parameter does not result in an increased data output bandwidth of the session, or when retransmission time-out occurs at a certain frequency or higher, the change in the TCP parameter is stopped for preventing a congestion on the network.

[0081]

Also, when the number of transmissible bytes of the session is smaller than another previously set value, the value of the TCP control parameter for the session is changed in order to reduce the data delivery through the TCP control to a bandwidth based on the assignment weight of packet scheduler 16.

[0082]

The former case involves an increase in the congestion window by a larger width during non-congestion or a reduction in a congestion window reducing rate during congestion, whereas the latter case involves processing reverse to the former case.

[0083]

Also, in the session relay apparatus according to this embodiment, weight changing unit 161 in packet scheduler 16 dynamically changes the assignment weight in accordance with a bandwidth currently available for transmission. Specifically, assignment weight changing unit 161 determines that data cannot be delivered in a bandwidth which has been set based on the assignment weight, if the number of transmissible bytes of the

session is not larger than the previously set value that is larger than the assignment weight, and temporarily reduces the assignment weight.

[0084]

Subsequently, when the number of transmissible bytes becomes smaller than the previously set other value, assignment weight changing unit 161 increases the assignment weight in succession to the original value. Also, assignment weight changing unit 161 calculates a bandwidth in which the session can transmit, based on the congestion window for the session notified from session relay unit 12-1 - 12-N and a measured value of round-trip propagation delay time.

[0085]

Further, if this calculated value differs from a bandwidth set value based on the assignment weight by a certain threshold or more, assignment weight changing unit 161 temporarily changes the assignment weight such that the bandwidth set value based on the assignment weight is equal to the calculated value mentioned above.

(Third Embodiment)

Fig. 9 is a block diagram illustrating the configuration of a session relay apparatus according to a third embodiment of the present invention. In Fig. 9, the session relay apparatus according to the third embodiment of the present invention is similar in configuration to the session relay apparatus according to the first embodiment of the present invention illustrated in Fig. 1, except that

reception rate control units 151-1 - 151-N (reception rate control units 151-2 - 151-N are not shown) are provided in session relay units 15-1 - 15-N, and the same components are designated the same reference numerals. Also, the operation of the same components is similar to the first embodiment of the present invention. Reception rate control units 151-1 - 151-N controls a reception rate from transmission terminal 3.

[0086]

While the operation of the session relay apparatus according to the third embodiment of the present invention will be described with reference to Fig. 9, description will be herein given only of aspects different from the first embodiment of the present invention described above.

[0087]

In the session relay apparatus according to this embodiment, when free capacity is exhausted in transmission buffer 122-1 - 122-N, reception session processing unit 123-1 - 123-N notifies transmission terminal 3 that the advertisement window size is zero, and transmission terminal 3 stops transmitting data in response thereto.

[0088]

In the first embodiment of the present invention, even if free capacity is available in transmission buffer 122-1 - 122-N at a later time, ACK cannot be delivered for resuming the transmission until transmission terminal 3 delivers a packet for a window test.

[0089]

In this embodiment, on the other hand, when packet scheduler 13 instructs delivery of a packet , reception rate control unit 151-1 tests the free capacity of transmission buffer 122-1 - 122-N after the packet delivery, and when this is one or MSS or more, an ACK packet is generated for transmission terminal 3 in order to prompt the same for rapid resumption of transmission.

[0090]

Also, when free capacity or an average thereof increases to a previously set certain value or more in transmission buffer 122-1 - 122-N, reception rate control unit 151-1 - 151-N instructs transmission terminal 3 to reduce the transmission bandwidth in order to prevent the exhaustion of the free capacity in transmission buffer 122-1 - 122-N.

[0091]

This can be achieved, for example, by replacing an ACK packet returned to the transmission side with a duplicate ACK, discarding received packets, setting an ECN (Explicit Congestion Notification) bit in the ACK packet, delaying the ACK packet, temporarily rewriting the advertisement window of the ACK packet to be small, and the like.

(Fourth Embodiment)

Fig. 10 is a block diagram illustrating the configuration of a session relay apparatus according to a fourth embodiment of the present invention. In Fig. 10, the session relay apparatus according to the fourth embodiment of the present invention is similar in configuration to the session relay apparatus according to

the third embodiment of the present invention except that session relay units 17-1 - 17-N are provided with transmission iSCSI (internet Small Computer System Interface) control units 171-1 - 171-N (transmission iSCSI control units 171-2 - 171-N are not shown) and reception iSCSI control units 172-1 - 172-N (reception iSCSI control units 172-2 - 172-N are not shown), and the same components are designated the same reference numerals. Also, the operation of the same components is similar to the session relay apparatus according to the third embodiment of the present invention.

[0092]

Transmission iSCSI control unit 171-1 - 171-N reflects congestion control information for the iSCSI layer to the transmission rate to reception terminal 2. Reception iSCSI control unit 172-1 - 172-N reflects congestion control information for the iSCSI layer to the reception rate from transmission terminal 3.

[0093]

The operation of the session relay apparatus according to the fourth embodiment of the present invention will be described with reference to Fig. 10. Here, a description will be given only of different aspects between the fourth embodiment of the present invention and the third embodiment of the present invention.

[0094]

On the iSCSI layer, an R2T (Ready-To-Transfer) packet is transmitted from reception terminal 2 to transmission terminal 3 for notifying the amount of data which can be received by reception

terminal 2 in order to conduct transfer control between transmission terminal 3 and reception terminal 2.

[0095]

Then, in this embodiment, when transmission iSCSI control unit 171-1 - 171-N receives the R2T packet from reception terminal 2, this R2T packet is not sent to reception terminal 2, but the sequence number of the last one of data delivered upon receipt of the R2T packet is stored in the session relay apparatus according to this embodiment.

[0096]

Transmission iSCSI control unit 171-1 - 171-N adds the amount of transmissible data of the R2T packet received this time to the sequence number stored the last time, and notifies packet scheduler 13 of the sum as the amount of data which can be transmitted on the iSCSI layer.

[0097]

Packet scheduler 13 defines the number of transmission waiting bytes as a minimum value of the amount of data permitted by the TCP layer for transmission and the amount of data permitted by the iSCSI layer. Specifically, as shown in Fig. 11, the number of transmission waiting bytes = min(sequence number of the last unit of sequentially received data + 1, advertisement window indicated by reception terminal, congestion window of the session, sequence number when R2T packet was received the last time + the amount of transmissible data of R2T packet received this time) - a transmitted sequence number.

[0098]

When packet scheduler 13 instructs delivery of a packet , reception iSCSI control unit 172-1 - 172-N tests transmission buffer 122-1 - 122-N for a free capacity after the packet delivery, and transmits the free capacity of the transmission buffer to transmission terminal 3 as a R2T packet if this is equal to or more than a previously defined certain value or more than this value.

[0099]

As described above, congestion control is not conducted for the iSCSI layer between transmission terminal 3 and reception terminal 2, but congestion control for the iSCSI layer is relayed by the session relay apparatus, thereby making it possible to achieve an improvement in the throughput on the iSCSI layer as well, similar to that when a TCP session is relayed.

(Fifth Embodiment)

Fig. 12 is a block diagram illustrating the configuration of a session relay apparatus according to a fifth embodiment of the present invention. In Fig. 12, the session relay apparatus according to the fifth embodiment of the present invention comprises session identification unit 11, session transmission units 41-1 - 41-N, session reception units 42-1 - 42-N, packet scheduler 13, and delivery control unit 14.

[0100]

Session identification unit 11 determines a session to which an incoming packet belongs, session transmission unit 41-1 - 41-N performs transmission processing for data in a session to a

reception terminal, and session reception unit 42-1 - 42-N performs reception processing for data from a session to the reception terminal.

[0101]

Packet scheduler 13 controls the delivery of packets from each session transmission unit 41-1 - 41-N. Delivery control unit 14 delivers packets from each session transmission unit 41-1 - 41-N based on instructions from packet scheduler 13.

[0102]

Session transmission units 41-1 - 41-N comprise transmission session processing units 411-1 - 411-N (transmission session processing units 411-2 - 411-N are not shown), transmission data generation units 412-1 - 412-N (transmission data generation units 412-2 - 412-N are not shown), and transmission buffers 413-1 - 413-N (transmission buffers 413-2 - 413-N are not shown).

[0103]

Transmission data generation unit 413-1 - 413-N stores transmission data from an application program in transmission buffer 413-1 - 413-N. Transmission buffer 5-6-1 temporarily stores data to be transmitted. Transmission session processing unit 411-1 - 411-N processes a session for transmitting data to the reception terminal.

[0104]

Session reception units 42-1 - 42-N comprise reception session processing units 421-1 - 421-N (reception session

processing units 421-2 - 421-N are not shown), reception buffers 422-1 - 422-N (reception buffers 422-2 - 422-N are not shown), and received data processing units 423-1 - 423-N.

[0105]

Reception session processing unit 421-1 - 421-N perform reception processing for data from the reception terminal, while reception buffer 422-1 - 422-N temporarily store received data. Received data processing unit 423-1 - 423-N pass received data from reception buffer 5-14-1 to an application.

[0106]

Generally, since bidirectional communications are made between a transmission terminal and a reception terminal in a TCP session, one of session transmission units 41-1 - 41-N and one of session reception units 42-1 - 42-N are used for the set of a transmission terminal and a reception terminal. In this embodiment, the session relay apparatus additionally serves as a transmission terminal or a reception terminal.

[0107]

Fig. 13 is a block diagram illustrating the flow of data between the session relay apparatus (transmission terminal) and session relay apparatus (reception terminal) according to the fifth embodiment of the present invention. In Fig. 13, when data is transmitted from session relay apparatus (transmission terminal) 4-2 to session relay apparatus (reception terminal) 4-1, a data packet delivered from session transmission unit 41-1-2 of session relay apparatus (transmission terminal 4-2) undergoes reception

processing in session reception unit 42-1-1 of session relay apparatus (reception terminal) 4-1. As a result, the generated ACK packet is returned to session transmission unit 41-1-2 of session relay apparatus (transmission terminal) 4-2.

[0108]

On the other hand, when data is sent from session relay apparatus (reception terminal) 4-1 to session relay apparatus (transmission terminal) 4-2, the data packet delivered from session transmission unit 41-1-1 in session relay apparatus (reception terminal) 4-1 undergoes reception processing in session reception unit 42-1-2 of session relay apparatus (transmission terminal) 4-2. As a result, the generated ACK packet is returned to session transmission unit 41-1-1 of session relay apparatus (reception terminal) 4-1.

[0109]

Next, the operation of the fifth embodiment of the present invention will be described with reference to Fig. 12. First, a description will be given of data transfer from the transmission terminal to the reception terminal.

[0110]

Transmission data generated by an application program is written into transmission buffer 413-1 - 413-N by transmission data generation unit 412-1 - 412-N. Transmission session processing unit 411-1 - 411-N processes data which has been transmitted from transmission buffer 413-1 - 413-N to the reception terminal. Since this processing is similar to the aforementioned first embodiment of

the present invention, a description thereon is omitted.

[0111]

Here, for calculating the number of transmission waiting bytes, packet scheduler 13 employs the sequence number of the last unit of data received from the application program instead of the sequence number of the last one of sequentially received data.

[0112]

Next, description will be given of data transfer from the reception terminal to the transmission terminal. Reception session processing unit 421-1 - 42-1N performs reception processing for data transmitted from the reception terminal, and stores data which can be correctly received in transmission buffer 422-1 - 422-N.

[0113]

This reception processing is similar to the aforementioned first embodiment of the present invention except that the data is stored in the reception buffer instead of in the transmission buffer, so that a description thereon is omitted. The data written into reception buffer 422-1 - 422-N is retrieved by reception data processing unit 423-1 - 423-N, and passed to the application program.

[0114]

Thus, in the present invention, the creation of congestion control information alone is performed on each of a plurality of layers in regard to congestion control processing for the layers, and packet delivery control processing is concentrated in a scheduler on the IP layer, thereby making it possible to realize session relay

processing at higher speeds.

[0115]

Also, the present invention can realize the session relay processing at higher speeds by concentrating reception buffers and transmission buffers for a plurality of layers in the transmission buffer for the IP layer to eliminate data movements between buffers.